Dynastat, Inc.
Research and Testing Services for the Voice Communications Community
2704 Rio Grande, Suite 4
Austin, TX 78705 USA
Phone: +1-512-476-4797 FAX: +1-512-472-2883
Contacts: Alan D. Sharpley (sharpley@dynastat.com)
Ira L. Panzer (ilpanzer@dynastat.com)

DYNASTAT – METHODS AND PROCEDURES FOR AM AUDIO TESTING

Prepared for:

Dr. Ellyn Sheffield
iBiquity Digital Corp.
20 Independence Blvd.
Warren, NJ

31 December 2001

INTRODUCTION

Dynastat has performed the data collection for the subjective testing of the In-Band On-Channel (IBOC) hybrid system developed by iBiquity Digital Corporation. This phase of the testing effort has involved performance of IBOC in AM mode. Dynastat has completed data collection for six experiments in this phase of the test program. The following sections describe the methods and procedures of the subjective testing effort conducted at Dynastat's laboratory.

BACKGROUND

Dynastat, Inc., Austin, Texas was formed in 1974 by Dr. William Voiers and his colleagues Alan Sharpley and Ira Panzer. Dynastat has a long history of involvement with the evaluation of voice communication systems using subjective testing methods. Over the past two decades this has consisted of the development and implementation of methods for measuring speech intelligibility, speech quality, and speaker recognizability. Dynastat personnel have developed the Diagnostic Rhyme Test (DRT), the Diagnostic Acceptability Measure (DAM), and the Diagnostic Speaker Recognizablity Test (DSRT). The DRT is one of the ANSI standards for measuring speech intelligibility (ANSI S3.2-1989) and the DAM has become a *de facto* standard for measuring speech quality at the Department of Defense. In addition, Dynastat has implemented most other methods currently in use for assessing performance of speech coding systems. For assessing speech intelligibility, these include the Modified Rhyme Test (MRT) and the Phonetically Balanced Word Test (PB), the other two ANSI standards along with the DRT. For measuring speech quality, Dynastat has also implemented all the International Telecommunication Union (ITU) standards (P.800), including the Absolute Category Rating (ACR) method which yields the Mean Opinion Score (MOS), the Degradation Category Rating (DCR) method from which the Degradation Mean Opinion Score (DMOS) is derived and the Comparison Category Rating (CCR) method which yields the CMOS.

It is Dynastat's policy to work with all our customers in determining their testing needs and to be equipped to provide virtually all subjective testing methods. Over the past two decades Dynastat has contracted with various international and national standard bodies to conduct subjective listener tests as an independent testing laboratory. These groups have included the International Telecommunication Union-Telecommunications Sector (ITU-T), the European Telecommunications Standards Institute (ETSI), and the Telecommunication Industry Association (TIA) as well as both Third-Generation Partnership Projects, 3-GPP and 3-GPP2. Dynastat is active in all of these standards groups providing expertise on test design, test implementation, test evaluation, and data analyses. Dynastat was approached by iBiquity to perform the evaluations for the NRSC effort. Except for the creation of the new test lab as explained in the body of this report, all systems were in place to undertake this effort.

SUBJECTIVE TESTING FACILITIES

Dynastat designed and built an AM Audio Testing Laboratory for this phase of the testing project. The laboratory is comprised of three isolated and sound-treated rooms with a measured ambient noise level < 38dBA. The room dimensions are approximately 10 ft. x 10 ft. with 8 ft. ceilings, a volume of approximately 800 cubic feet. Each room contains a listening station that includes a chair, an HP Vectra VL400 PC, a high-quality Lucid DA9624 digital to analog converter, and two Tannoy Nearfield Audio Monitors mounted on 36" speaker stands – the speaker is approximately head-high for a typical seated adult. The placement of the chair and the two audio monitors was maintained such that the listener's head and the two monitors formed an equilateral triangle with approximately 48" sides. Figure 1 shows one of the sound-treated rooms with placement of listener and listening station. The PC's, A/D converters, and monitors were provided to Dynastat by iBiquity. Sound samples were stored on the hard-disk of each PC and are presented to the listeners under program control using a software package developed by iBiquity. The software also displays the appropriate rating scale(s) on the monitor and collects and stores the listener's responses. Each listening station is independent and self-contained and requires no experimenter control or interaction once the listener has started an experiment. Dynastat's AM Audio Testing Laboratory includes the three rooms/listening stations with the individual PC's networked to a server PC used for loading audio files and compiling listener responses.

Fig. 1 One of the listening stations contained in Dynastat's AM Audio Testing Laboratory.

LISTENER SAMPLE

The sample of listeners for the NRSC subjective experiments was stratified both for listener gender and age-group. For each experiment listeners were recruited to represent approximately equal representation in eight categories: four *Age-Groups* (16-24, 25-32, 33-42, 43-50) for each *Gender* (male, female). In general, each experiment required Dynastat to deliver the subjective data from 40 *qualified* listeners, where *qualification* was based on performance on an initial screening test developed by iBiquity and a post-hoc screening test designed to eliminate obvious outliers. Most of the listeners for the AM phase were recruited from a pool of 400 listeners who had participated in an earlier FM phase of testing for iBiquity. All of these listeners had passed both the initial and post hoc screening in the FM phase of testing. The remaining listeners were recruited from a pool of more than 2000 listeners contained in Dynastat's subjective testing database. That database is a continually evolving and expanding pool of listeners that Dynastat has maintained for use in subjective evaluation of speech-coding and voice-communications systems. Membership in Dynastat's subjective database is largely dictated by guidelines specified by ITU-T[1] and other standardization bodies. This latter group of listeners (i.e., those who had not participated in the FM phase of testing) were provided with additional training and tested using the initial screening test.

[1] ITU-T Recommendation P.800, Methods for subjective determination of transmission quality, Aug., 1996.

PROCEDURES

Upon arrival at Dynastat, listeners completed a brief biographical data-sheet and received verbal instructions on the specific tasks to be performed in the experiment. Exhibit A shows the instructions that were read to listeners. Each listener was assigned a unique eight-character listener ID (i.e., **Eeeesai**) coded for experiment (**Eeee**), gender (**s** = 1 for male, 2 for female), age-group (**a** = 1 for 16-24, 2 for 25-32, etc.), and individual (i.e., individual within the category, **i** = 1, 2, etc.). For example, the ID "*A01a112*" would identify the 2nd individual listener who was a male listener in age-group 16-25 participating in experiment A01a. The test administrator entered the listener's ID and biographical information into an Excel *Participant* file specific to the experiment. The overall duration of each experiment was approximately 1.5 hours and typically included a training phase and a testing phase consisting of one or more test sessions. For those listeners who had to be screened, there was an additional half-hour test session that involved pre-screening training and the screening test. The overall test duration is within the maximum testing time recommended by the ITU-T's recommendation P.800.

Training and Testing in the Screening Phase

Those listeners who had not been screened in the previous FM phase of testing were required to go through an initial training and screening session. During the training phase of this session listeners were presented a range of audio impairments typical of those involved in the testing phase of the experiment. The training phase was developed and provided to Dynastat by iBiquity and was used to expose and familiarize the listeners to the variety and range of conditions they were likely to hear in the subsequent screening and testing phases. The impairments presented in the training phase ranged from subtle to extreme and served to train the listeners to listen carefully for potential impairments in the audio samples. The training materials were presented at a group listening station in the Audio Laboratory equipped with a Rane HC-6 distribution amplifier, which allowed the test administrator, and up to four listeners to hear the training materials over the Sennheiser HD-600 headphones. There were seven training samples, each involving multiple cuts. In each training sample, the first cut was a "clean" cut followed by two or more "impaired" cuts of the same materials. Each listener was asked if he could tell the difference between the cuts, i.e., if he could hear the impairments. The sample was replayed until all listeners acknowledged that they could hear the impairments. The experimenter asked only if the listeners could "hear the differences between the cuts." The experimenter never discussed the specific types of impairments involved in the training samples or how the listeners should judge or value those impairments.

Immediately after the training, listeners participated in a pre-test screening to ensure that they were able to reliably distinguish between "clean" and "impaired" samples. The listener's task in the screening phase was a "Reference-A-B" comparison in which the listener was required to decide which of two "test" samples (A or B) was the same as the reference sample. In each trial one of the test samples was the same as the clean or unimpaired reference sample and the other sample was an impaired sample. Figure 2 shows the PC response display that was used for the screening task. Playback of samples was under the individual listener's control, but the screening software required him to listen to all three samples, reference and two test samples, before the response options were available. Listeners were free to replay any or all of the three samples until they were ready to enter their response and proceed to the next trial. The screening phase consisted of one practice trial and ten test trials. Listeners were provided no feedback on the "correctness" of their responses during the screening test. After completion of the screening phase, the listeners exited their booth for a short rest-break during which the test administrator scored their screening responses. Listeners were not informed of their specific performance in the screening phase, but depending on their score, were placed in one of three categories. If a listener scored 50% or less (i.e., 5 of 10 correct or at the "chance" level) he was paid a partial fee for his participation and was not allowed to proceed to the test phase of the experiment. If a listener scored 60% or 70% he was allowed to proceed to the test phase but his data was not used in the final set of ratings delivered to iBiquity (i.e., his data was disqualified based on screening test performance). Listeners who scored 80-100% proceeded to the test phase as a "qualified" listener and their rating data was used in "post-hoc" screening designed to provide the most reliable data possible. Description of the "post-hoc" data screening is provided in a later section.

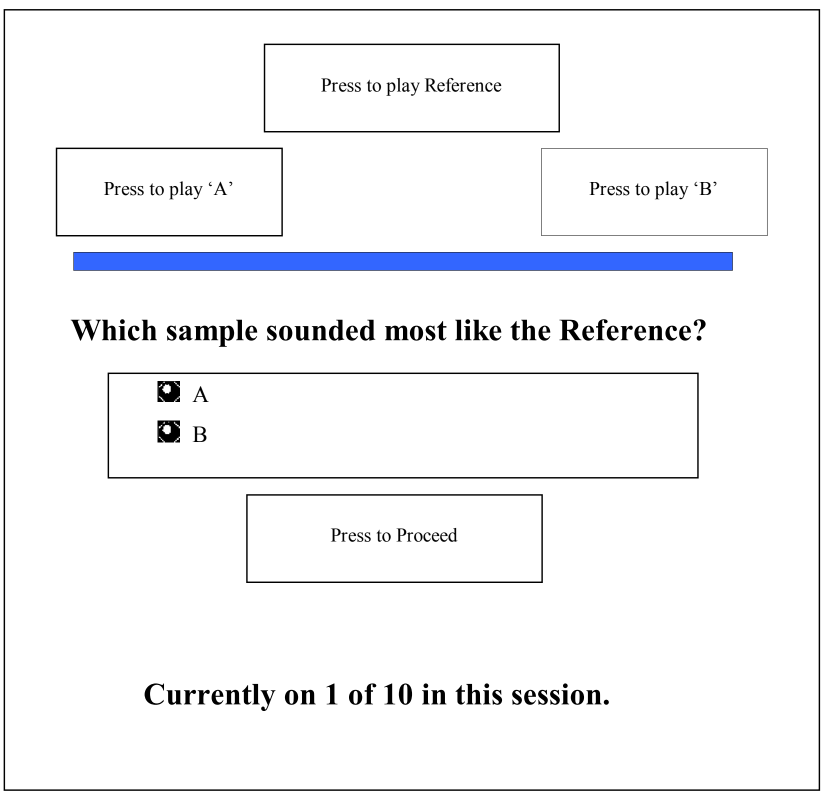

Fig.2 Response display for the Ref/A/B task in the screening phase.

AM Training

Listeners participated in a short training session (i.e., seven trials) to familiarize them with the kinds and degrees of impairments that they would experience in the AM audio testing session. The training was self-paced and was conducted at the listening stations described above. For each trial, the listeners task was to listen to two audio samples, a clean sample and an impaired sample, and confirm that (s)he could perceive the difference between the two samples.

AM Testing

Table I presents a summary of the six experiments that Dynastat has conducted for the AM testing effort. The Absolute Category Rating (ACR) method was the subjective evaluation tool in all of these. The ACR method yields the Mean Opinion Score (MOS), a measure of overall audio quality. The ACR requires the listener to judge the quality of an audio sample using a five category rating scale where: Excellent=5, Good=4, Fair=3, Poor=2, and Bad=1. The category judgments are reported as a measure of overall audio quality, an MOS, on a scale of 1 to 5. A response display for the ACR testing task is shown in Fig. 3. The listener controlled playback of the audio samples but on each trial he could enter his response only after listening to the entire sample. Typically, the testing phase consisted of two practice trials followed by approximately 200 test trials. The listener could adjust the playback volume during the practice trials. The playback volume set by the listener during the practice trials was then maintained throughout the remainder of the experiment. Test trials were grouped into sessions of approximately 50 trials each, separated by rest-breaks. During the rest-breaks listeners were required to leave the listening room.

Table I. Summary of AM experiments conducted by Dynastat.

Exp.	Test Methodology	# Audio Samples	# Listeners Retained	# Listeners Excluded	Min. FoM	Min. FoM Index
A01	ACR	222	40	5	.708	.855
A02	ACR	190	40	6	.518	.772
A03a-c	ACR	289	60	10	.706	.850

A04a-e	**ACR**	**262**	**50**	**2**	**.739**	**.872**
A05	**ACR**	**213**	**40**	**2**	**.705**	**.879**
A06	**ACR**	**174**	**40**	**4**	**.635**	**.851**

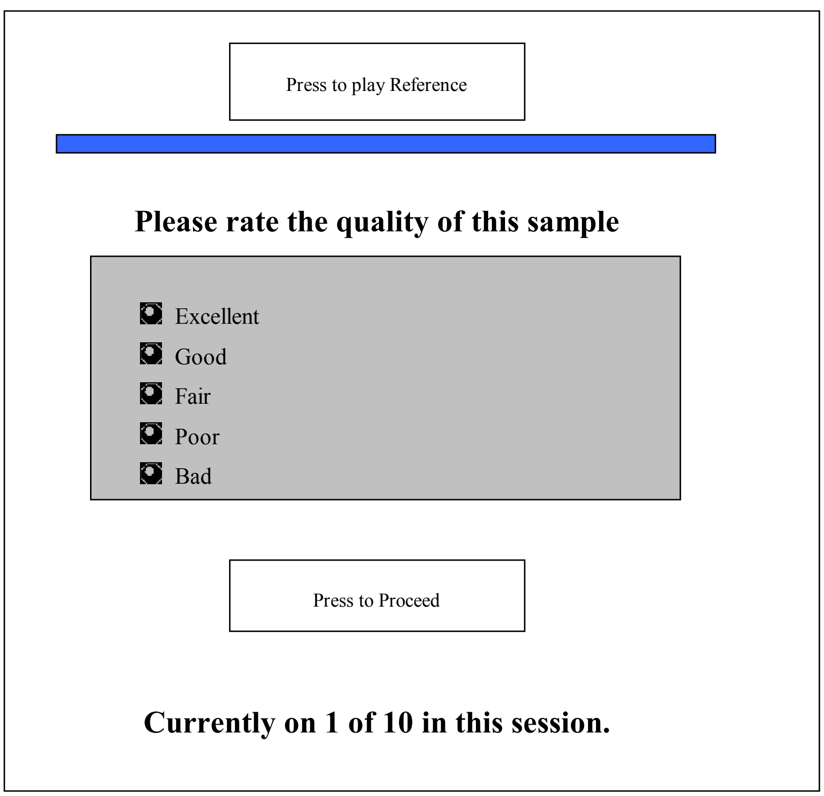

Fig. 3. Response display for the ACR task in the testing phase.

AUDIO MATERIALS

For each experiment approximately 200 processed audio samples were supplied to Dynastat by iBiquity. The audio materials were delivered to Dynastat via Internet FTP. The files were provided in digital format (44.1 kHz, 16 bit linear WAV). The digital files were loaded onto the hard-disk of the server PC and then distributed to the hard-disks of the individual PC's through a local area network.

AUDIO FILE PRESENTATION AND DATA COLLECTION

For each experiment Dynastat prepared an Excel file that controlled the audio file presentation and data collection software. During this process the audio files were loaded and verified, file order randomizations were created, and the overall layout of the experiment was established (i.e., number of test sessions, number of trials per session, and number of rest-breaks). The iBiquity software package automatically accumulated the listener responses into an experiment specific Excel *Response* spreadsheet. Once a test session had been initiated, the iBiquity software required no input from the test administrator. File presentation and data collection were controlled by the interaction of the listener and the software.

POST-HOC DATA ANALYSIS AND LISTENER SCREENING

At the conclusion of the data collection for an experiment, the total set of listener data (i.e., the experiment *Response* Excel file) was subjected to a post-hoc analysis to ensure the validity and the reliability of the data for each individual listener. For each experiment a "Figure of Merit" (FoM) was calculated for each listener participating in the experiment. The FOM was the "coefficient of correlation" between the individual listener's vector of ratings and the vector containing the average ratings for the remainder of the listeners involved in the experiment. Thirty years of experience with subjective rating data has shown this FoM to be a valuable screening measure to remove clear "outliers" from the rating data (i.e., listeners who either can't or won't perform the rating task). A practical lower threshold of 0.70[2] for the FoM was generally used to classify listeners as "outliers" and remove their data from the experiment. The last two columns in Table I show, for each experiment, the minimum value of FoM for the listeners that were retained in the final set of data delivered to iBiquity as well as the FoM index for the experiment. After eliminating listeners from the data set on the bases of pre-test and post-hoc screening, it was sometime necessary to remove additional listeners in order to satisfy the sample

[2] Since the FoM is based on a correlation coefficient it is subject both to the amount of variation in the rating data as well as the range of that data. The criterion value of .70 was arbitrarily chosen on the basis of empirical evidence and past experience in subjective testing efforts. For individual experiments the criterion value was adjusted according to the variation and range of the observed data.

stratification requirements. In case where one or more *qualified* listeners had to be removed from a specific gender/age-group category, listeners were randomly selected for deletion.

DATA DELIVERY

For each experiment Dynastat compiled and delivered two Excel worksheets to iBiquity. The *Participant* worksheet contained the biographical and ID information for the 40 listeners contained in the final data set. The *Response* worksheet contained the raw response data for those listeners.

Exhibit A - Instructions for ACR Audio Rating Experiments

Overview

Welcome to this audio testing session. Today, you will be participating in a listening experiment which should last about two and a half hours. You will be listening to music and speech samples over headphones. We are studying how various radios sound under different transmission conditions. There are three parts to this study. The first part is training, where you will listen to the music you will be encountering in your tests. The second part is a discrimination test. The third part is an opinion test.

Training Task

In the training session, you will hear a variety of sound samples. These sound samples include typical transmission "impairments" you might hear during the discrimination and opinion tests. These impairments should be noticeable. During the course of each sample you will hear varying degrees of the "impairment". You will indicate to the administrator if differences are heard.

Discrimination Task

In the discrimination task we will be testing your ability to hear different impairments. In this task your job is to decide which of two samples (A or B) is most nearly the same as the reference sample. The response display is shown in Fig. 1. To begin click on the box labeled "Press to Play Reference". The complete reference sample will be played. Similarly, you will click on "Press to Play A" and "Press to Play B" to play these complete samples. The program will not let you enter a response until you have heard all three samples completely. After listening to the complete Reference, A, and B samples you can enter your response to the question "Which sample sounded most like the reference?". After indicating your response click on the box labeled "Press to proceed". If you would like to play any of the samples again, you can press the appropriate box and do so as much as needed until you have made your decision. Once you have indicated your response and clicked on the "Press to proceed" you will be ready to start your next trial. During the course of your practice trial for this task you can set the volume level my moving the slider box. Once this level is set it cannot be changed for the rest of the session.

The discrimination session will consist of one practice trial and 10 test trials. When you complete the task open the door and proceed to the waiting room for a 10-minute break. During the break the administrator will score your data and let you know if you passed the test. If you passed the test then you are eligible to participate in the opinion test. If you did not pass you will be paid $20 for your efforts.

Opinion Task – The ACR-MOS Test

In this part of the experiment we are evaluating systems that might be used for the radio transmission of sound samples. You are going to hear a number of recorded samples and rating how good you think they sound.

On each trial a single sample will be presented. Each sample will consist of a 10-15 second music or voice passage. Please listen to the complete sample, then indicate your opinion of the overall sound quality of the sample using the following 5-point scale: Exellent, Good, Fair, Poor, Bad. Figure 3 shows the response display.

This task is different from the discrimination task. There is no stated reference against which to compare the samples you are hearing. You simply hear a passage and then make a rating. You will have to use an internal reference to judge "the goodness" of the sample. By that we mean, when you are listening to a particular sample, think about how a very good radio station would sound in your car and over your home radio. Judge the sample in relation to your memory of those two references.

Many things go into a quality rating. You'll be listening for impairments as well as the overall aesthetic quality. By aesthetic we mean beauty, musicality, character, sound quality, etc. Try to judge each sample in an overall sense. This is especially hard to do if a big impairment happens to occur at the end of the sample. So, before you rate each sample, take a few seconds to think about the entire sample you just heard. In that way, it won't be just your last impression that carries the most weight.

The experiment will involve four test sessions separated by short rest periods. In the first session you will have a practice block of 2 trials to familiarize you with the rating task and adjust your listening volume. The practice block will be followed by 4 test sessions of 50 trials each. If you have any questions, please feel free to ask the test administrator.

Please do not discuss your opinions with any other listeners participating in the experiment. Thank you in advance for your participation.